



# Multiplicative Bias Corrected Nonparametric Smoothers

Nicolas Hengartner, Eric Matzner-Løber, Laurent Rouvière, Thomas Burr

## ► To cite this version:

Nicolas Hengartner, Eric Matzner-Løber, Laurent Rouvière, Thomas Burr. Multiplicative Bias Corrected Nonparametric Smoothers. Nonparametric statistics, Springer proceedings in Mathematics and Statistics, 3rd ISNPS, Avignon, 2016., 2018, 10.1007/978-3-319-96941-1\_3 . hal-00408696v3

**HAL Id: hal-00408696**

**<https://hal.science/hal-00408696v3>**

Submitted on 28 Feb 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# MULTIPLICATIVE BIAS CORRECTED NONPARAMETRIC SMOOTHERS

## Abstract

The paper presents a multiplicative bias reduction estimator for nonparametric regression. The approach consists to apply a multiplicative bias correction to an oversmooth pilot estimator. In Burr et al. [2010], this method has been tested to estimate energy spectra. For such data set, it was observed that the method allows to decrease bias with negligible increase in variance. In this paper, we study the asymptotic properties of the resulting estimate and prove that this estimate has zero asymptotic bias and the same asymptotic variance as the local linear estimate. Simulations show that our asymptotic results are available for modest sample sizes.

**Index terms:** Nonparametric regression, bias reduction, local linear estimate.

## 1 Introduction

In nonparametric regression, the bias-variance tradeoff of linear smoothers such as kernel-based regression smoothers, wavelet based smoother or spline smoothers, is generally governed by a user-supplied parameter. This parameter is often called the bandwidth. As an example, assuming that the regression function  $m$  is twice continuously differentiable, the local linear

smoother with bandwidth  $h$  and kernel  $K$  has conditional bias

$$\frac{h^2}{2}m''(x) \int u^2 K(u) \, du + o_p(h^2)$$

and conditional variance

$$\frac{1}{nh} \frac{\sigma^2(x)}{f(x)} \int K^2(u) \, du + o_p\left(\frac{1}{nh}\right)$$

where  $f$  stands for the density of the (one-dimensional) explanatory variable  $X$  and  $\sigma^2(x)$  is the conditional variance of the response variable given  $X = x$ . See for example the book of Fan and Gijbels [1996]. Since the bias increases with the second order derivative of the regression function, the local linear smoother tends to under-estimate in the peaks and over-estimate in the valleys of the regression function. See for example Simonoff [1996], Wand and Jones [1995], Scott [1992]. This behavior in peaks and valleys causes some trouble in some practical applications, such as for the estimation of energy spectrum.

The decay of radioactive isotopes often generates gamma particles whose energy can be measured using specialized detectors. Typically, these detectors count the number of particles in various energy bins over short time intervals such as one to ten minutes. This enables estimation of the energy distribution of the emitted particles, which is called the energy spectrum. For low or medium resolution detectors, the spectrum is typically composed of multiple broad peaks whose location and area characterize the radio-isotope. That is we not only want the locations of the peaks, but also the shape and amplitudes in peak regions.

Because the actual bin counts are noisy, and the energy spectrum is fairly smooth, it has been proposed to estimate the energy spectrum using non-parametric smoothing techniques (Sullivan et al. [2006], Gang et al. [2004]).

However, the bias of the nonparametric smoothers degrades isotope identification performance for any algorithm that includes peak area or ratios of areas (Casson et al. [2006]) and motivates studying methods to reduce bias.

All nonparametric smoothing methods are generally biased. There are many approaches to reducing the bias, but most of them do so at the cost of an increase in the variance of the estimator. For example, one may choose to undersmooth the energy spectrum. Undersmoothing will reduce the bias but will have a tendency of generating spurious peaks. One can also use higher order smoothers, such as local polynomial smoother with a polynomial of order larger than one. While again this will lead to a smaller bias, the smoother will have a larger variance. Another approach is to start with a pilot smoother and to estimate its bias by smoothing the residuals (Di Marzio and Taylor [2008]). Subtracting the estimated bias from the smoother produces a regression smoother with smaller bias and larger variance. For the estimation of an energy spectrum, the additive bias correction and the higher order smoothers have the unfortunate side effect of possibly generating a non-positive estimate.

An attractive alternative to the linear bias correction is the multiplicative bias correction pioneered by Linton and Nielsen [1994]. Because the multiplicative correction does not alter the sign of the regression function, this type of correction is particularly well suited for adjusting non-negative regression functions. Jones et al. [1995] showed that if the true regression function has four continuous derivatives, then the multiplicative bias reduction is operationally equivalent to using an order four kernel. And while this does remove the bias, it also increases the variance.

Many authors have extended the work of Jones et al. [1995]. Glad [1998a,b]

proposes to use a parametrically guided local linear smoother and Nadaraya-Watson smoother by starting with a parametric pilot. This approach is extended to a more general framework which includes both multiplicative and additive bias correction by Martins-Filho et al. [2008] (see also Mishra et al. [2010] for an extension to time series conditional variance estimation). For multiplicative bias correction in density estimation, we refer the reader to the recent works of Gustafsson et al. [2009], Hagmann and Scaillet [2007] and Hirukawa [2010].

Although the bias-variance tradeoff for nonparametric smoothers is always present in finite samples, it is possible to construct smoothers whose *asymptotic bias* converges to zero while keeping the same asymptotic variance. Hengartner and Matzner-Løber [2009] has exhibited a nonparametric density estimator based on multiplicative bias correction with that property, and have shown in simulations that their estimator also enjoyed good finite sample properties. Burr et al. [2010] present such an estimator for nonparametric regression to estimate energy spectra. They illustrate the benefits of this approach on real and simulated spectra. The goal of this paper is to study the asymptotic properties of this estimator. It is worth pointing out that these properties have already been studied by Linton and Nielsen [1994] for fixed design and further by Jones et al. [1995]. We emphasize that there are two major differences between our work and that of Jones et al. [1995].

- First, we do not add regularity assumptions on the target regression function. In particular, we do not assume that the regression function has four continuous derivatives as in Jones et al. [1995].
- Second, we show that the multiplicative bias reduction procedure performs a bias reduction with no cost to the asymptotic variance. It is

exactly the same as the asymptotic variance of the local linear estimate.

We provide another asymptotic behavior under less restrictive assumptions than in Jones et al. [1995]. Moreover our results and proofs are completely different from the above referenced works.

This paper is organized as follows. Section 2 introduces the notation and defines the estimator. Section 3 gives the asymptotic behavior of the proposed estimator. A brief simulation study on finite sample comparison is presented in Section 4. The interested reader is referred to the Appendix where we have gathered the technical proofs.

## 2 Preliminaries

### 2.1 Notation.

Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be  $n$  independent copies of the pair of random variables  $(X, Y)$  with values in  $\mathbb{R} \times \mathbb{R}$ . We suppose that the explanatory variable  $X$  has probability density  $f$  and model the dependence of the univariate response variable  $Y$  to the explanatory variable  $X$  through the nonparametric regression model

$$Y = m(X) + \varepsilon. \quad (1)$$

We assume that the regression function  $m(\cdot)$  is smooth and that the disturbance  $\varepsilon$  is a mean zero random variable with finite variance  $\sigma^2$  that is independent of the covariate  $X$ . Consider the linear smoothers for the regression function  $m(x)$  which we can write as

$$\hat{m}(x) = \sum_{j=1}^n \omega_j(x; h) Y_j,$$

where the weight function  $\omega_j(x; h)$  depends on a tuning parameter  $h$ , that we think of as the bandwidth.

If the weight functions are such that  $\sum_{j=1}^n \omega_j(x; h) = 1$  and  $\sum_{j=1}^n \omega_j(x; h)^2 = (nh)^{-1}\tau^2$ , and if the disturbances satisfy the Lindberg-Feller condition, then the linear smoother obeys the central limit theorem

$$\sqrt{nh} \left( \hat{m}(x) - \sum_{j=1}^n w_j(x; h) m(X_j) \right) \rightarrow \mathcal{N}(0, \tau^2). \quad (2)$$

We can use (2) to construct asymptotic pointwise confidence intervals for the unknown regression function  $m(x)$ . But unless the limit of the scaled bias

$$b(x) = \lim_{n \rightarrow \infty} \sqrt{nh} \left( \sum_{j=1}^n w_j(x; h) m(X_j) - m(x) \right),$$

which we call the asymptotic bias, is zero, the confidence interval

$$\left[ \hat{m}(x) - Z_{1-\alpha/2} \sqrt{nh} \tau, \hat{m}(x) + Z_{1-\alpha/2} \sqrt{nh} \tau \right]$$

will not cover asymptotically the true regression function  $m(x)$  at the nominal  $1 - \alpha$  level. The construction of valid pointwise  $1 - \alpha$  confidence intervals for regression smoothers is another motivation for developing estimators with zero asymptotic bias.

## 2.2 Multiplicative bias reduction

Here we present a framework for multiplicative bias reduction. Given a pilot smoother

$$\tilde{m}_n(x) = \sum_{j=1}^n \omega_j(x; h_0) Y_j,$$

the ratio

$$V_j = \frac{Y_j}{\tilde{m}_n(X_j)}$$

is a noisy estimate of  $m(X_j)/\tilde{m}_n(X_j)$ , the inverse relative estimation error of the smoother  $\tilde{m}_n$  at each of the observations. Smoothing  $V_j$  by

$$\hat{\alpha}_n(x) = \sum_{j=1}^n \omega_j(x; h_1) V_j$$

yields an estimate for the inverse of the relative estimation error which can be used as a multiplicative correction of the pilot smoother. This leads to the (nonlinear) smoother

$$\hat{m}_n(x) = \hat{\alpha}_n(x) \tilde{m}_n(x). \quad (3)$$

The estimator (3) was studied for fixed design by Linton and Nielsen [1994] and further studied by Jones et al. [1995]. In both cases, they assumed that the regression function had four continuous derivatives, and show an improvement in the convergence rate of the corrected estimator. Glad [1998a,b] proposed to use a parametrically guided local linear smoother and Nadaraya-Watson smoother by starting with a parametric pilot. She shows that the resulting estimates improve on the local polynomial estimate as soon as the pilot captures some of the features of the regression function.

### 3 Theoretical Analysis of Multiplicative Bias Reduction

In this section, we show that the multiplicative smoother has smaller bias with essentially no cost to the variance, assuming only two derivatives of the regression function. While the derivation of our results are for local linear smoothers, the technique used in the proofs can be easily adapted for other linear smoothers, and the conclusions remain essentially unchanged.



### 3.1 Assumptions

We make the following assumptions:

1. The regression function is bounded and strictly positive, that is,  $b \geq m(x) \geq a > 0$  for all  $x$ .
2. The regression function is twice continuously differentiable everywhere.
3. The density of the covariate is strictly positive on the interior of its support in the sense that  $f(x) \geq b(\mathcal{K}) > 0$  over every compact  $\mathcal{K}$  contained in the support of  $f$ .
4.  $\varepsilon$  has finite fourth moments and has a symmetric distribution around zero.
5. Given a symmetric probability density  $K(\cdot)$ , consider the weights  $\omega_j(x; h)$  associated to the local linear smoother. That is, denote by  $K_h(\cdot) = K(\cdot/h)/h$  the scaled kernel by the bandwidth  $h$  and define for  $k = 0, 1, 2, 3$  the sums

$$S_k(x) \equiv S_k(x; h) = \sum_{j=1}^n (X_j - x)^k K_h(X_j - x).$$

Then

$$\omega_j(x; h) = \frac{S_2(x; h) - (X_j - x)S_1(x; h)}{S_2(x; h)S_0(x; h) - S_1^2(x; h)} K_h(X_j - x).$$

We set

$$\omega_{0j}(x) = \omega_j(x; h_0) \quad \text{and} \quad \omega_{1j}(x) = \omega_j(x; h_1).$$

6. The bandwidths  $h_0$  and  $h_1$  are such that

$$h_0 \rightarrow 0, \quad h_1 \rightarrow 0, \quad nh_0 \rightarrow \infty, \quad nh_1^3 \rightarrow \infty, \quad \frac{h_1}{h_0} \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty.$$

The positivity assumption (assumption 1) on  $m(x)$  is classical when we perform a multiplicative bias correction. It allows to avoid that the terms  $Y_j/\tilde{m}_n(X_j)$  blows up. Of course, the regression function might cross the  $x$ -axis. For such a situation, Glad [1998a] proposes to shift all response data  $Y_i$  a distance  $a$ , so that the new regression function  $m(x) + a$  does not any more intersect with the  $x$ -axis. Such a method can also be performed here. Assumptions 2–4 are standard to obtain rate of convergence for nonparametric estimators. Assumption 5 means that we conduct the theory for the local linear estimate. The results can be generalized to other linear smoothers. Assumption 6 is not restrictive since it is satisfied for a wide range of values of  $h_0$  and  $h_1$ .

### 3.2 A technical aside

The proof of the main results rests on establishing a stochastic approximation of estimator (3) in which each term can be directly analyzed.

**Proposition 3.1.** *We have*

$$\hat{m}_n(x) = \mu_n(x) + \sum_{j=1}^n \omega_{1j}(x) A_j(x) + \sum_{j=1}^n \omega_{1j}(x) B_j(x) + \sum_{j=1}^n \omega_{1j}(x) \xi_j,$$

where  $\mu_n(x)$ , conditionally on  $X_1, \dots, X_n$  is a deterministic function,  $A_j$ ,  $B_j$  and  $\xi_j$  are random variables. Under condition  $nh_0 \rightarrow \infty$ , the remainder  $\xi_j$  converges to 0 in probability and we have

$$\hat{m}_n(x) = \mu_n(x) + \sum_{j=1}^n \omega_{1j}(x) A_j(x) + \sum_{j=1}^n \omega_{1j}(x) B_j(x) + O_P\left(\frac{1}{nh_0}\right).$$

**Remark 3.1.** A technical difficulty arises because even though  $\xi_j$  may be small in probability, its expectation may not be small. We resolve this prob-

lem by showing that we only need to modify  $\xi_j$  on a set of vanishingly small probability to guarantee that its expectation is also small.

**Definition 3.1.** Given a sequence of real numbers  $a_n$ , say that a sequence of random variables  $\xi_n = o_p(a_n)$  if for all fixed  $t > 0$ ,

$$\limsup_{n \rightarrow \infty} \mathbb{P}[|\xi_n| > ta_n] = 0.$$

We will need the following Lemma.

**Lemma 3.1.** *If  $\xi_n = o_p(a_n)$ , then there exists a sequence of random variables  $\xi_n^*$  such that*

$$\limsup_{n \rightarrow \infty} \mathbb{P}[\xi_n^* \neq \xi_n] = 0 \quad \text{and} \quad \mathbb{E}[\xi_n^*] = o(a_n).$$

*We shall use the following notation*

$$\mathbb{E}_\star[\xi_n] = \mathbb{E}[\xi_n^*].$$

### 3.3 Main results

We deduce from Proposition 3.1 and Lemma 3.1 the following Theorem.

**Theorem 3.1.** *Under the assumptions (1)-(6), the estimator  $\widehat{m}_n$  satisfies:*

$$\mathbb{E}_\star(\widehat{m}_n(x) | X_1, \dots, X_n) = \mu_n(x) + O_p\left(\frac{1}{n\sqrt{h_0 h_1}}\right) + O_p\left(\frac{1}{nh_0}\right)$$

*and*

$$\mathbb{V}_\star(\widehat{m}_n(x) | X_1, \dots, X_n) = \sigma^2 \sum_{j=1}^n w_{1j}^2(x) + O_p\left(\frac{1}{nh_0}\right) + o_p\left(\frac{1}{nh_1}\right).$$

We deduce from Theorem 3.1 that if the bandwidth  $h_0$  of the pilot estimator converges to zero much slower than  $h_1$ , then  $\widehat{m}_n$  has exactly the same asymptotic variance as the local linear smoother of the original data with bandwidth  $h_1$ . However, for finite samples, the two step local linear smoother can have a slightly larger variance depending on the choice of  $h_0$ . For the bias term, a limited Taylor expansion of  $\mu_n(x)$  leads to the following result.

**Theorem 3.2.** *Under the assumptions (1)-(6), the estimator  $\widehat{m}_n$  satisfies:*

$$\mathbb{E}_\star(\widehat{m}_n(x)|X_1, \dots, X_n) = m(x) + o_p(h_1^2).$$

It is worth pointing out that we only suppose that the regression function is twice continuously differentiable. We do not add smoothness assumptions. For a study of the local linear estimate in the presence of jumps in the derivative, we refer the reader to Desmet and Gijbels [2009]. Combining Theorem 3.1 and Theorem 3.2, we conclude that the multiplicative adjustment performs a bias reduction on the pilot estimator without increasing the asymptotic variance. The asymptotic behavior of the bandwidths  $h_0$  and  $h_1$  is constrained by assumption 6. However, it is easily seen that this assumption is satisfied for a large set of values of  $h_0$  and  $h_1$ . For example, the choice  $h_1 = c_1 n^{-1/5}$  and  $h_0 = c_0 n^{-\alpha}$  for  $0 < \alpha < 1/5$  leads to

$$\mathbb{E}_\star(\widehat{m}_n(x)|X_1, \dots, X_n) - m(x) = o_p(n^{-2/5})$$

and

$$\mathbb{V}_\star(\widehat{m}_n(x)|X_1, \dots, X_n) = O_p(n^{-4/5}).$$

**Remark 3.2.** Estimators with bandwidths of order  $O(n^{-\alpha})$  for  $0 < \alpha < 1/5$  are oversmoothing the true regression function, and as a result, they have biases that are of larger order of magnitude than their standard deviations.

We conclude that the multiplicative adjustment performs a bias reduction on the pilot estimator.

## 4 Numerical examples

While the amount of the bias reduction depends on the curvature of the regression function, a decrease is expected (asymptotically) everywhere, and this, at no cost to the variance. The simulation study in this section shows that this asymptotic behavior emerges already at modest sample sizes.

### 4.1 Local study

To illustrate numerically the possible reduction in the bias and associated increase of the variance achieved by the multiplicative bias correction, consider estimating the regression function

$$m(x) = 5 + 3|x|^{5/2} + x^2 + 4\cos(10x)$$

at  $x = 0$  (see Figure 1). The local linear smoother tends to under-estimate the regression function at their maximum, and hence, this example will provide a good example. Furthermore, because the second derivative of this regression function is continuous but not differentiable at the origin, the results previously obtained by Linton and Nielsen [1994] do not apply.

The data are simulated according to the model

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \dots, 100,$$

where  $\varepsilon_i$  are independent  $\mathcal{N}(0, 1)$  variables. We first consider the local linear estimate with a Gaussian kernel function and we study its performances over

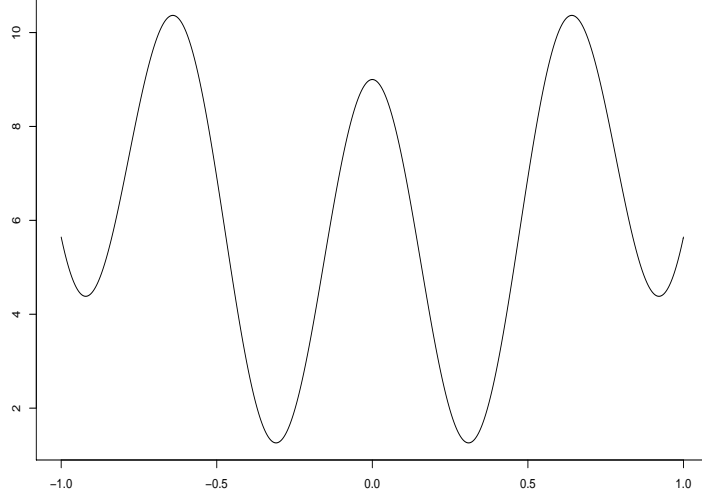


Figure 1: The regression function to be estimated.

a grid of bandwidths  $\mathcal{H} = [0.005, 0.1]$ . For the new estimate, the theory recommends to start with an oversmooth pilot estimate. In this regard, we take  $h_0 = 0.1$  and study the performance of the multiplicative bias corrected estimate for  $h_1 \in \mathcal{H}_1 = [0.005, 0.12]$ . To explore the stability of our two stages estimator with respect to  $h_0$ , we also consider the choice  $h_0 = 0.02$ . For such a choice, the pilot estimate clearly undersmooths the regression function.

Bias and variance of each estimate are calculated at  $x = 0$ . To do this, we compute the value of each estimate at  $x = 0$  for 200 samples  $(X_i, Y_i), i = 1, \dots, 100$ . The same design  $X_i, i = 1, \dots, 100$  is used for each sample. It is generated according to a uniform distribution over  $[-1, 1]$ . The bias at point  $x = 0$  is estimated by subtracting  $m(0)$  at the mean value of the estimate at  $x = 0$  (the mean value is computed over the 200 replications). Similarly we

estimate the variance at  $x = 0$  by the variance of the values of the estimate at this point. Figure 2 presents squared bias, variance and mean square error of each estimate for different values of bandwidths  $h$  for the local linear smoother and  $h_1$  for our estimate.

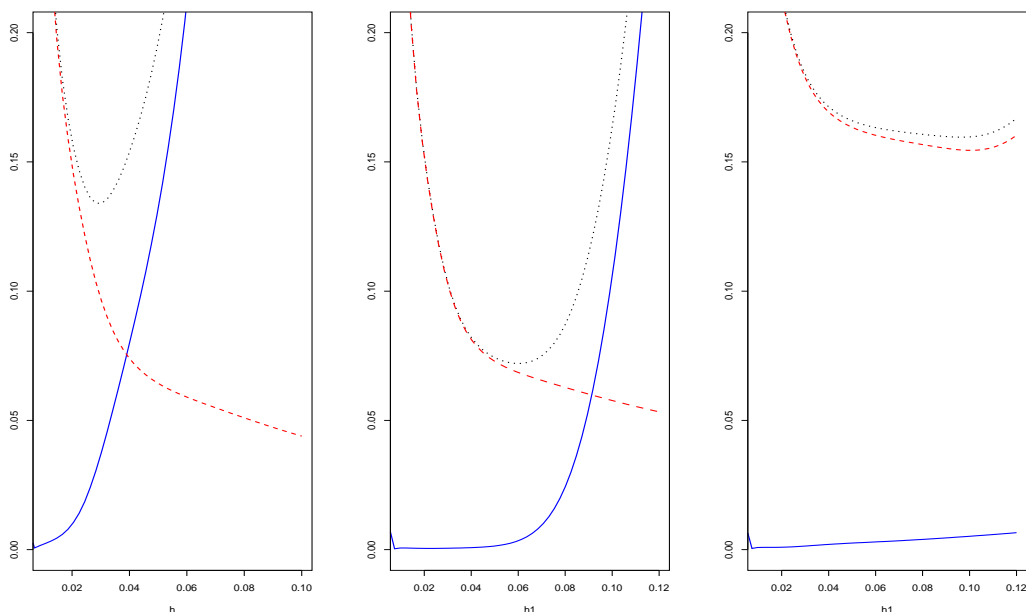


Figure 2: Mean square error (dotted line), squared bias (solid line) and variance (dashed line) of the local linear estimate (left) and multiplicative bias corrected estimate with  $h_0 = 0.1$  (center) and  $h_0 = 0.02$  (right) at point  $x = 0$ .

The first conclusion is that the corrected estimate has smaller bias than the local linear estimate provided the pilot estimate oversmooths the regression function. Small values of  $h_0$  clearly undersmooth the regression function, whatever the choice of  $h_1$ . Moreover, it is worth pointing out that our procedure does not significantly increase the variance. Even if Theorem 3.1 and Theorem 3.2 provide asymptotic results, our simulations show that the asymptotic behavior of our estimate emerges already at modest sample size.

Finally, due to the bias reduction, we note that our procedure also reduces the optimal mean square error (see Table 1).

	MSE	Bias <sup>2</sup>	Variance
LLE	0.134	0.038	0.096
MBCE	0.072	0.003	0.069

Table 1: Optimal mean square error (MSE) for the local linear estimate (LLE) and the multiplicative bias corrected estimate (MBCE) with  $h_0 = 0.1$  at point  $x = 0$ .

We conclude our local study with a comparison between our estimate and the estimate proposed by Glad [1998a]. To do this, we compute the multiplicative bias corrected estimate using three parametric starts:

- first the guide is chosen correctly and belong to the true parametric family:

$$\tilde{m}_n^1(x) = \hat{\beta}_0 + \hat{\beta}_1|x|^{5/2} + \hat{\beta}_2x^2 + \hat{\beta}_3 \cos(10x);$$

- second, we consider a linear parametric guide (which is obviously wrong):

$$\tilde{m}_n^2(x) = \hat{\beta}_0 + \hat{\beta}_1x;$$

- finally, we use a more reasonable guide, not correct, but that can reflect some a priori idea on the regression function

$$\tilde{m}_n^3(x) = \hat{\beta}_0 + \hat{\beta}_1x + \hat{\beta}_2x^2 + \dots + \hat{\beta}_8x^8.$$

All the estimates  $\hat{\beta}_j$  stands for the classical least square estimates.

The multiplicative bias correction is performed on these parametric starts using the local linear estimate. The performance of the resulting estimates



is measured over a grid of bandwidths  $\mathcal{H}_2 = [0.005; 0.4]$ . Bias and variance of each estimate are still estimated at  $x = 0$ . We keep the same setting as above and all the results are averaged over the same 200 replications. We display in Table 2 the optimal MSE calculated over the grid  $\mathcal{H}_2$ .

	MSE	Bias <sup>2</sup>	Variance
start $\tilde{m}_n^1$	0.060	0.000	0.060
start $\tilde{m}_n^2$	0.134	0.038	0.096
start $\tilde{m}_n^3$	0.095	0.021	0.074

Table 2: Optimal mean square error for the multiplicative bias corrected estimates with parametric starts  $\tilde{m}_n^j$ ,  $j = 1, 2, 3$ .

As expected, we first observe that the performance clearly depends on the choice of the parametric start. Table 1 and table 2 show that (in term of MSE) the estimate studied in this paper is better than the corrected estimated with parametric start  $\tilde{m}_n^2$  and  $\tilde{m}_n^3$ . Unsurprisingly, the best performance are obtained with the parametric guide  $\tilde{m}_n^1$  (which belongs to the true model). In practice, when one has no or few a priori information on the target regression function, the method proposed in the present paper is preferable.

## 4.2 Global study

This paper does not conduct any theory to select the two bandwidths  $h_0$  and  $h_1$  in an optimal way. If automatic procedures are needed, they can be obtained by adjusting traditional automatic selection procedures for the classical nonparametric estimators (see Burr et al. [2010]). In this part, we propose to use leave-one-out cross validation to choose both  $h_0$  and  $h_1$ .

We then compare the performance of the selected estimate with the local polynomial estimate in term of integrated square error.

Hurvich et al. [1998] report a comprehensive numerical study that compares standard smoothing methods on various test functions. Here, we take the same setting to compare the local linear estimate with its multiplicative bias corrected smoother. In each of the examples, we take the Gaussian kernel  $K(x) = \exp(-x^2/2)/\sqrt{2\pi}$ . We use the following regression functions (see Figure 3):

$$(1) \quad m_1(x) = \sin(5\pi x)$$

$$(2) \quad m_2(x) = \sin(15\pi x)$$

$$(3) \quad m_3(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$$

$$(4) \quad m_4(x) = 0.3 \exp[-64(x - .25)^2] + 0.7 \exp[-256(x - .75)^2]$$

and we take a Gaussian error distribution with standard deviation  $\sigma_j = 0.25\|m_j\|_2$  for  $j = 1, \dots, 4$ .

We use a cross validation device to select both  $h_0$  and  $h_1$ . This selection procedure involves solving minimization problem that necessitate a search over a finite grid  $\mathcal{H}$  of bandwidths  $h_0$  and  $h_1$ . Formally, given  $\mathcal{H}$ , we choose  $\hat{h}_0$  and  $\hat{h}_1$  such as

$$(\hat{h}_0, \hat{h}_1) = \underset{(h_0, h_1) \in \mathcal{H} \times \mathcal{H}}{\operatorname{argmin}} \quad \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{m}_n^i(X_i))^2.$$

Here  $\hat{m}_n^i$  stands for the corrected local polynomial estimate after deleted the  $i$ th observation. To assess the quality of the selected estimate, we compare its performances with the local polynomial estimate for which the bandwidth is again selected by leave-one-out cross validation. The performance of an estimator  $\hat{m}$  is measured by the integrated square error

$$ISE(\hat{m}) = \int_0^1 (m(x) - \hat{m}(x))^2 dx,$$

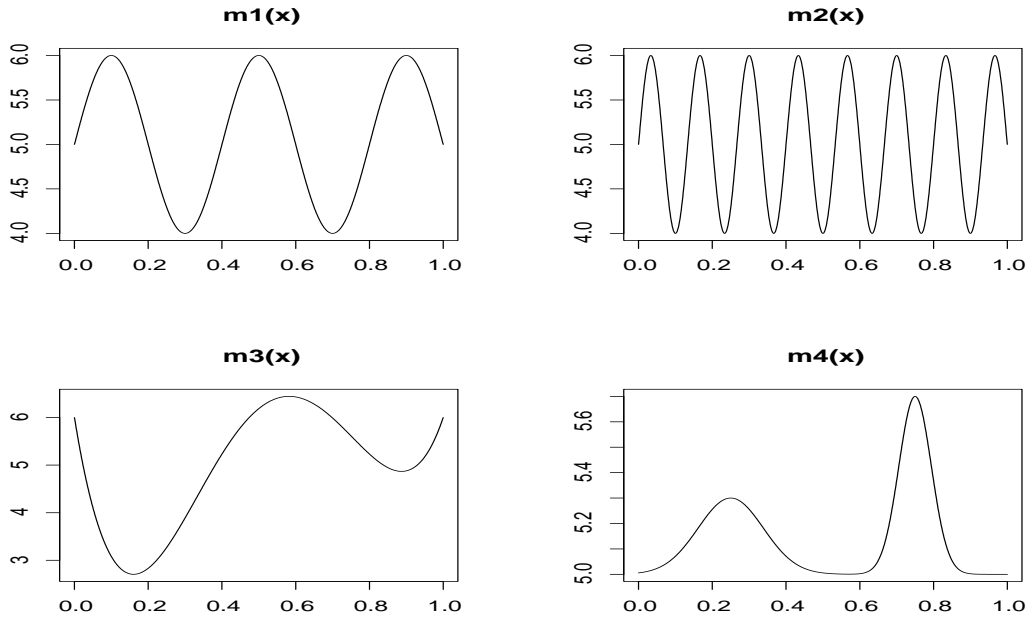


Figure 3: Regression functions to be estimated.

and to avoid the boundary effects, the design  $X_1, \dots, X_n$  is generated according to a uniform distribution over  $[-0.2, 1.2]$ .

Table 3 presents the median over 100 replications of

- the selected bandwidths;
- the integrated square error;
- the integrated square error of the local linear estimate divided by the integrated square error of the corrected estimate ( $R_{ISE}$ ).

Figure 4 displays the boxplots of the integrated square error for each estimate.

	LLE		MBCE			
	$h$	ISE ( $\times 100$ )	$h_0$	$h_1$	ISE ( $\times 100$ )	$R_{ISE}$
$m_1$	0.023	0.920	0.041	0.032	0.727	1.191
$m_2$	0.011	5.967	0.027	0.012	4.968	1.205
$m_3$	0.029	2.063	0.071	0.054	1.139	1.648
$m_4$	0.018	0.087	0.033	0.023	0.076	1.147

Table 3: Median over 100 replications of the selected bandwidths and of the integrated square error of the selected estimates. LLE and MBCE stands for local linear estimate and multiplicative bias corrected estimate.

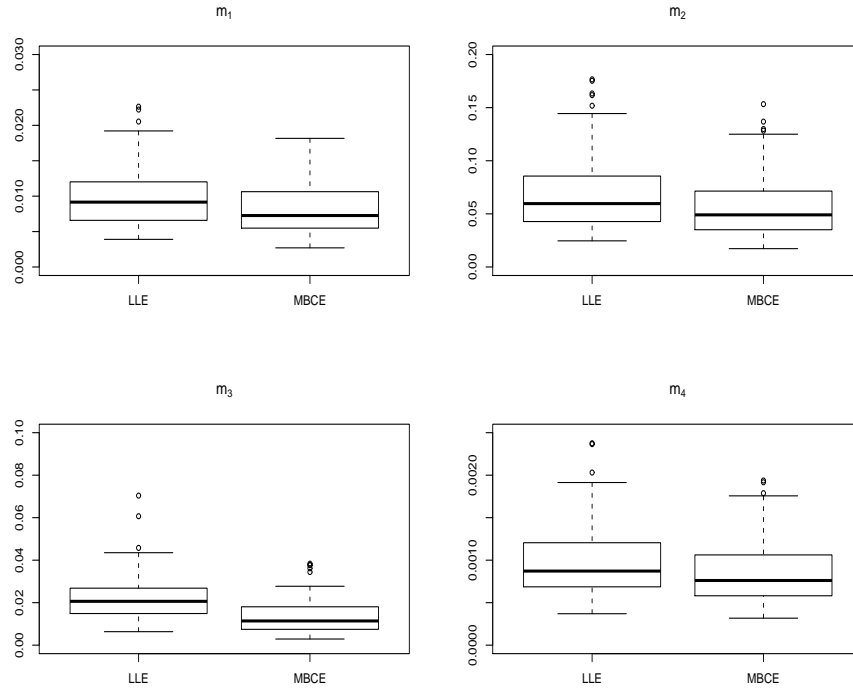


Figure 4: Boxplot of the integrated square error over the 100 replications.

We obtain significant ISE reduction for the four models. As predicted by

Theorem 3.1, the data-driven procedure selects  $h_0$  bigger than  $h$ : the pilot estimate is oversmoothing the true regression function. Of course, selecting both  $h_0$  and  $h_1$  is time consuming and can appear as the price to be paid to improve the local linear smoother.

Figure 5 presents, for the regression function  $m_1$  with  $n = 100$  and 100 iterations, different estimators on a grid of points. In lines is the true regression function which is unknown. For every point on a fixed grid, we plot, side by side, the mean over 100 replications of our estimator at that point (left side) and on the right side of that point the mean over 100 replications of the local polynomial estimator. Leave-one-out cross validation is applied to select the bandwidths  $h_0$  and  $h_1$  for our estimator and the bandwidth  $h$  for the local polynomial estimator. We add also the interquartile interval in order to see the fluctuations of the different estimators.

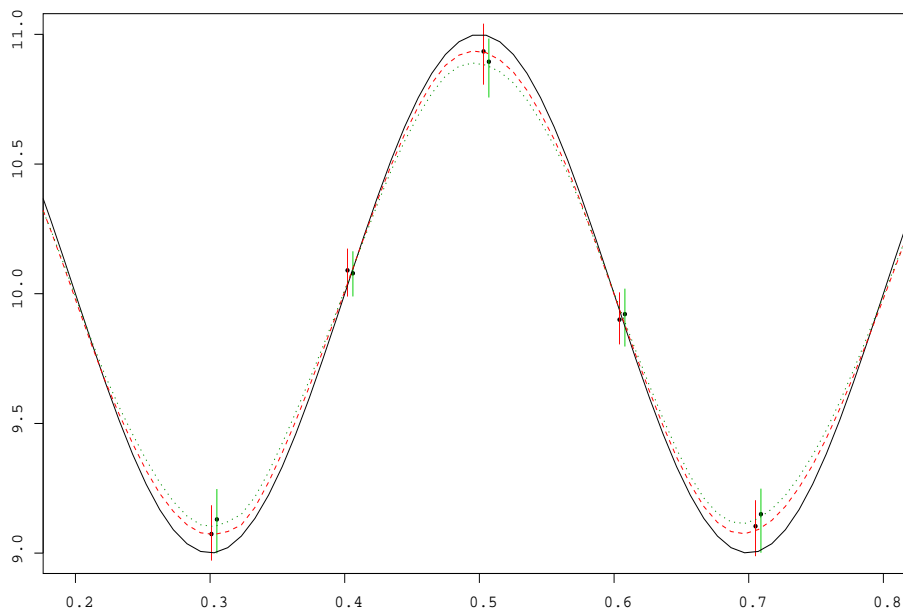


Figure 5: The solid curve represents the true regression function, our estimator is in dashed line and local linear smoother is dotted.

In this example, our estimator reduces the bias by increasing the peak and decreasing the valleys. Moreover, the interquartile intervals look similar for both estimator, as predicted by the theory.

## 5 Proofs

This section is devoted the technical proofs.

### 5.1 Proof of Proposition 3.1

Write the bias corrected estimator

$$\hat{m}_n(x) = \sum_{j=1}^n \omega_{1j}(x) \frac{\tilde{m}_n(x)}{\tilde{m}_n(X_j)} Y_j = \sum_{j=1}^n \omega_{1j}(x) R_j(x) Y_j,$$

and let us approximate the quantity  $R_j(x)$ . Define

$$\bar{m}_n(x) = \sum_{j=1}^n \omega_{0j}(x) m(X_j) = \mathbb{E}(\tilde{m}_n(x) | X_1, \dots, X_n),$$

and observe that

$$\begin{aligned} R_j(x) &= \frac{\tilde{m}_n(x)}{\tilde{m}_n(X_j)} \\ &= \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)} \times \left( 1 + \frac{\tilde{m}_n(x) - \bar{m}_n(x)}{\bar{m}_n(x)} \right) \times \left( 1 + \frac{\tilde{m}_n(X_j) - \bar{m}_n(X_j)}{\bar{m}_n(X_j)} \right)^{-1} \\ &= \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)} \times [1 + \Delta_n(x)] \times \frac{1}{1 + \Delta_n(X_j)}, \end{aligned}$$

where

$$\Delta_n(x) = \frac{\tilde{m}_n(x) - \bar{m}_n(x)}{\bar{m}_n(x)} = \frac{\sum_{l \leq n} \omega_{0l}(x) \varepsilon_l}{\sum_{l \leq n} \omega_{0l}(x) m(X_l)}.$$

Write now  $R_j(x)$  as

$$R_j(x) = \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)} [1 + \Delta_n(x) - \Delta_n(X_j) + r_j(x, X_j)]$$

where  $r_j(x, X_j)$  is a random variable converging to 0 to be define latter on. Given the last expression and model (1), estimator (3) could be written as

$$\begin{aligned}
\hat{m}_n(x) &= \sum_{j=1}^n \omega_{1j}(x) R_j(x) Y_j \\
&= \sum_{j=1}^n \omega_{1j}(x) \frac{\tilde{m}_n(x)}{\tilde{m}_n(X_j)} m(X_j) \\
&\quad + \sum_{j=1}^n \omega_{1j}(x) \frac{\tilde{m}_n(x)}{\tilde{m}_n(X_j)} [\varepsilon_j + m(X_j) (\Delta_n(x) - \Delta_n(X_j))] \\
&\quad + \sum_{j=1}^n \omega_{1j}(x) \frac{\tilde{m}_n(x)}{\tilde{m}_n(X_j)} (\Delta_n(x) - \Delta_n(X_j)) \varepsilon_j \\
&\quad + \sum_{j=1}^n \omega_{1j}(x) \frac{\tilde{m}_n(x)}{\tilde{m}_n(X_j)} r_j(x, X_j) Y_j \\
&= \mu_n(x) + \sum_{j=1}^n \omega_{1j}(x) A_j(x) + \sum_{j=1}^n \omega_{1j}(x) B_j(x) + \sum_{j=1}^n \omega_{1j}(x) \xi_j.
\end{aligned}$$

which is the first part of the proposition. Under assumption set forth in Section 3.1, the pilot smoother  $\tilde{m}_n$  converges to the true regression function  $m(x)$ . Bickel and Rosenblatt [1973] show that this convergence is uniform over compact sets  $\mathcal{K}$  contained in the support of the density of the covariate  $X$ . As a result

$$\sup_{x \in \mathcal{K}} |\tilde{m}_n(x) - m(x)| \leq \frac{1}{2}.$$

So a limited expansion of  $(1 + u)^{-1}$  yields for  $x \in \mathcal{K}$

$$R_j(x) = \frac{\tilde{m}_n(x)}{\tilde{m}_n(X_j)} [1 + \Delta_n(x) - \Delta_n(X_j) + O_p(|\Delta_n(x)\Delta_n(X_j)| + \Delta_n^2(X_j))],$$

thus

$$\xi_j = O_p(|\Delta_n(x)\Delta_n(X_j)| + \Delta_n^2(X_j)).$$

Under the stated regularity assumptions, we deduce that

$$\xi_j = O_p\left(\frac{1}{nh_0}\right).$$

leading to the announced result. Proposition (3.1) is proved.

## 5.2 Proof of lemma (3.1)

By definition

$$\limsup_{n \rightarrow \infty} \mathbb{P}[|\xi_n| > ta_n] = 0$$

for all  $t > 0$ , so that a triangular array argument shows that there exists an increasing sequence  $m = m(k)$  such that

$$\mathbb{P}\left[|\xi_n| > \frac{a_n}{k}\right] \leq \frac{1}{k} \quad \text{for all } n \geq m(k).$$

For  $m(k) \leq n \leq m(k+1) - 1$ , define

$$\xi_n^* = \begin{cases} \xi_n & \text{if } |\xi_n| < k^{-1}a_n \\ 0 & \text{otherwise.} \end{cases}$$

It follows from the construction of  $\xi_n^*$  that for  $n \in (m(k), m(k+1) - 1)$ ,

$$\mathbb{P}[\xi_n \neq \xi_n^*] = \mathbb{P}[|\xi_n| > k^{-1}a_n] \leq \frac{1}{k},$$

which converges to zero as  $n$  goes to infinity. Finally set  $k(n) = \sup\{k : m(k) \leq n\}$ , we obtain

$$\mathbb{E}[|\xi|_n^*] \leq \frac{a_n}{k(n)} = o(a_n).$$

## 5.3 Proof of Theorem (3.1)

Recall that

$$\widehat{m}_n(x) = \mu_n(x) + \sum_{j=1}^n \omega_{1j}(x) A_j(x) + \sum_{j=1}^n \omega_{1j}(x) B_j(x) + O_P\left(\frac{1}{nh_0}\right).$$



Focus on the conditional bias, we get

$$\begin{aligned}\mathbb{E}(\mu_n(x)|X_1, \dots, X_n) &= \mu_n(x) \\ \mathbb{E}(A_j(x)|X_1, \dots, X_n) &= 0 \\ \mathbb{E}(B_j(x)|X_1, \dots, X_n) &= \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)} \sigma^2 \left( \frac{\omega_{0j}(x)}{\bar{m}_n(x)} - \frac{\omega_{0j}(X_j)}{\bar{m}_n(X_j)} \right).\end{aligned}$$

Since

$$\sum_{j=1}^n \omega_{1j}(x) \omega_{0j}(x) \leq \sqrt{\sum_{j=1}^n \omega_{1j}(x)^2} \sqrt{\sum_{j=1}^n \omega_{0j}(x)^2} = O_p \left( \frac{1}{n\sqrt{h_0 h_1}} \right),$$

we deduce that

$$\mathbb{E} \left( \sum_{j=1}^n \omega_{1j}(x) B_j(x) \middle| X_1, \dots, X_n \right) = O_p \left( \frac{1}{n\sqrt{h_0 h_1}} \right).$$

This proves the first part of the Theorem.

For the conditional variance, we use the following expansion of the two stages estimator

$$\widehat{m}_n(x) = \sum_{j=1}^n \omega_{1j}(x) \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)} Y_j (1 + [\Delta_n(x) - \Delta_n(X_j)]) + O_p \left( \frac{1}{nh_0} \right).$$

Using the fact that the residuals have four finite moments and have a symmetric distribution around 0, a moment's thought shows that

$$\mathbb{V}(Y_j [\Delta_n(x) - \Delta_n(X_j)] | X_1, \dots, X_n) = O_p \left( \frac{1}{nh_0} \right)$$

and

$$\text{Cov}(Y_j, Y_j [\Delta_n(x) - \Delta_n(X_j)] | X_1, \dots, X_n) = O_p \left( \frac{1}{nh_0} \right).$$

Hence

$$\begin{aligned}\mathbb{V}_\star(\widehat{m}_n(x)|X_1, \dots, X_n) &= \mathbb{V} \left( \sum_{j=1}^n \omega_{1j}(x) \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)} Y_j \middle| X_1, \dots, X_n \right) \\ &\quad + O_p \left( \frac{1}{nh_0} \right).\end{aligned}$$

Observe that the first term on the right hand side of this equality can be seen as the variance of the two stages estimator with a deterministic pilot estimator. It follows from Glad [1998a] that

$$\mathbb{V} \left( \sum_{j=1}^n \omega_{1j}(x) \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)} Y_j \middle| X_1, \dots, X_n \right) = \sigma^2 \sum_{j=1}^n \omega_{1j}^2(x) + o_p \left( \frac{1}{nh_1} \right),$$

which proves the theorem.

## 5.4 Proof of theorem (3.2)

Recall that

$$\mu_n(x) = \sum_{j \leq n} \omega_{1j}(x) \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)} m(X_j).$$

We consider the limited Taylor expansion of the ratio

$$\frac{m(X_j)}{\bar{m}_n(X_j)} = \frac{m(x)}{\bar{m}_n(x)} + (X_j - x) \left( \frac{m(x)}{\bar{m}_n(x)} \right)' + \frac{1}{2} (X_j - x)^2 \left( \frac{m(x)}{\bar{m}_n(x)} \right)'' (1 + o_p(1)),$$

then

$$\begin{aligned} \mu_n(x) &= \bar{m}_n(x) \left\{ \frac{m(x)}{\bar{m}_n(x)} \sum_{j=1}^n \omega_{1j}(x) + \left( \frac{m(x)}{\bar{m}_n(x)} \right)' \sum_{j=1}^n (X_j - x) \omega_{1j}(x) \right. \\ &\quad \left. + \frac{1}{2} \left( \frac{m(x)}{\bar{m}_n(x)} \right)'' \sum_{j=1}^n (X_j - x)^2 \omega_{1j}(x) (1 + o_p(1)) \right\}. \end{aligned}$$

It is easy to verify that

$$\begin{aligned} \Sigma_0(x; h_1) &= \sum_{j=1}^n \omega_{1j}(x) = 1, \\ \Sigma_1(x; h_1) &= \sum_{j=1}^n (X_j - x) \omega_{1j}(x) = 0 \\ \Sigma_2(x; h_1) &= \sum_{j=1}^n (X_j - x)^2 \omega_{1j}(x) = \frac{S_2^2(x; h_1) - S_3(x; h_1) S_1(x; h_1)}{S_2(x; h_1) S_0(x; h_1) - S_1^2(x; h_1)}. \end{aligned}$$

For random designs, we can further approximate (see, e.g., Wand and Jones [1995])

$$S_k(x, h_1) = \begin{cases} h^k \sigma_K^k f(x) + o_p(h^k) & \text{for } k \text{ even} \\ h^{k+1} \sigma_K^{k+1} f'(x) + o_p(h^{k+1}) & \text{for } k \text{ odd,} \end{cases}$$

where  $\sigma_K^k = \int u^k K(u) du$ . Therefore

$$\begin{aligned} \Sigma_2(x; h_1) &= h_1^2 \int u^2 K(u) du + o_p(h_1^2) \\ &= \sigma_K^2 h_1^2 + o_p(h_1^2), \end{aligned}$$

so that we can write  $\mu_n(x)$  as

$$\begin{aligned} \mu_n(x) &= \bar{m}_n(x) \left\{ \frac{m(x)}{\bar{m}_n(x)} + \frac{\sigma_K^2 h_1^2}{2} \left( \frac{m(x)}{\bar{m}_n(x)} \right)'' + o_p(h_1^2) \right\} \\ &= m(x) + \frac{\sigma_K^2 h_1^2}{2} \bar{m}_n(x) \left( \frac{m(x)}{\bar{m}_n(x)} \right)'' + o_p(h_1^2). \end{aligned}$$

Moreover

$$\begin{aligned} \left( \frac{m(x)}{\bar{m}_n(x)} \right)'' &= \frac{\bar{m}_n^2(x) m''(x)}{\bar{m}_n^3(x)} - 2 \frac{\bar{m}_n(x) \bar{m}_n'(x) m'(x)}{\bar{m}_n^3(x)} \\ &\quad - \frac{m(x) \bar{m}_n(x) \bar{m}_n''(x)}{\bar{m}_n^3(x)} + 2 \frac{m(x) (\bar{m}_n'(x))^2}{\bar{m}_n^3(x)} \end{aligned}$$

and applying the usual approximations, we conclude that

$$\left( \frac{m(x)}{\bar{m}_n(x)} \right)'' = o_p(1).$$

Putting all pieces together, we obtain

$$\mathbb{E}_\star(\hat{m}_n(x) | X_1, \dots, X_n) - m(x) = o_p(h_1^2) + O_p\left(\frac{1}{n\sqrt{h_0 h_1}}\right) + O_p\left(\frac{1}{nh_0}\right).$$

Since

$$nh_1^3 \longrightarrow \infty \quad \text{and} \quad \frac{h_1}{h_0} \longrightarrow 0,$$

we conclude that the bias is of order  $o_p(h_1^2)$ .

## References

- P. Bickel and M. Rosenblatt. On some global measures of the deviations of density function estimates. *The Annals of Statistics*, 1:1071–1095, 1973.
- T. Burr, N. Hengartner, E. Matzner-Løber, S. Myers, and L. Rouvière. Smoothing low resolution gamma spectra. *IEEE Transactions on Nuclear Science*, 57:2831–2840, 2010.
- W. Casson, C. Sullivan, J. Blackadar, R. Paternoster, J. Matzke, M. Rawool-Sullivan, and L. Atencio. *Nuclear Reachback Reference Manual*, chapter 6. LosAlamos National Laboratory, 2006.
- L. Desmet and I. Gijbels. Local linear fitting and improved estimation near peaks. *The Canadian Journal of Statistics*, 37:473–475, 2009.
- M. Di Marzio and C. Taylor. On boosting kernel regression. *Journal of Statistical Planning and Inference*, 138:2483–2498, 2008.
- J. Fan and I. Gijbels. *Local Polynomial Modeling and Its Application, Theory and Methodologies*. Chapman et Hall, New York, 1996.
- X. Gang, D. Li, Z. Benai, and Z. Jianshi. A nonlinear wavelet method for data smoothing of low-level gamma-ray spectra. *J. Nuclear Science and Technology*, 41:73–76, 2004.
- I. Glad. Parametrically guided non-parametric regression. *Scandinavian Journal of Statistics*, 25:649–668, 1998a.
- I. Glad. A note on unconditional properties of parametrically guided Nadaraya-Watson estimator. *Statistics and Probability Letters*, 37:101–108, 1998b.

- J. Gustafsson, M. Hagmann, J. Nielsen, and O. Scaillet. Local transformation kernel density estimation of loss distributions. *Journal of Business and Economic Statistics*, 27:161–175, 2009.
- M. Hagmann and O. Scaillet. Local multiplicative bias correction for asymmetric kernel density estimators. *Journal of Econometrics*, 141:213–249, 2007.
- N. Hengartner and E. Matzner-Løber. Asymptotic unbiased density estimators. *ESAIM*, 13:1–14, 2009.
- M. Hirukawa. Nonparametric multiplicative bias correction for kernel-type density estimation on the unit interval. *Computational Statistics and Data Analysis*, 54:473–495, 2010.
- C. Hurvich, G. Simonoff, and C. L. Tsai. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society*, 60:271–294, 1998.
- M. Jones, O. Linton, and J. Nielsen. A simple and effective bias reduction method for kernel density estimation. *Biometrika*, 82:327–338, 1995.
- O. Linton and J. P. Nielsen. A multiplicative bias reduction method for nonparametric regression. *Statistics and Probability Letters*, 19(181–187), 1994.
- C. Martins-Filho, S. Mishra, and A. Ullah. A class of improved parametrically guided nonparametric regression estimators. *Econometric Reviews*, pages 542–573, 2008.

- S. Mishra, L. Su, and A. Ullah. Semiparametric estimator of time series conditional variance. *Journal of Business and Economic Statistics*, 28: 256–274, 2010.
- D. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New-York, 1992.
- J. Simonoff. *Smoothing Methods in Statistics*. Springer, New York, 1996.
- C. Sullivan, M. Martinez, and S. Garner. Wavelet analysis of sodium iodide spectra. *IEEE transaction on nuclear science*, 53:2916–2922, 2006.
- M. Wand and M. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995.